The 2016 Conference on Computational Linguistics and Speech Processing
ROCLING 2016, pp. 311-324

基於深層類神經網路之音訊事件偵測系統

Deep Neural Networks for Audio Event Detection

陳智偉　Jhih-wei Chen
國立台北科技大學電子工程系
Department of Electronic Engineering
National Taipei University of Technology
t104368109@gmail.com

劉佳鑫　Chia-Hsin Liu
國立台北科技大學電子工程系
Department of Electronic Engineering
National Taipei University of Technology
Jeff81227@gmail.com

廖元甫　Yuan-Fu Liao
國立台北科技大學電子工程系
Department of Electronic Engineering
National Taipei University of Technology
yfliao@mail.ntut.edu.tw

摘要

現實生活中常有許多聲音事件會一起發生，而聲音會重疊在一起，使得傳統(Gaussian Mixture Model ,GMM)方法很難準確辨認這些重疊的聲音事件。因此，本文提出以深層類神經網絡(Deep Neural Network, DNN)來檢測這些互相干擾的聲音事件，並據此參加 Detection and Classification of Acoustic Scenes and Events 2016 (DCASE2016) 比賽，DCASE2016 評比提供的音訊資料，內有兩種場景，包括居家與戶外，共有 18 種含有背景的聲音事件。實驗結果顯示使用 DNN 與傳統 GMM 比較，其場景偵測錯誤率可從 0.91 降至 0.86、F1 分數並從 23.4%提升到 26.8%。此外針對室內環境的音訊事件偵測，錯誤率可從 1.06 降至 0.86，F1 分數並從 8.9%提升到 27.7%。最後在戶外環境的音訊偵測情境中，錯誤率可從 1.03 降至 0.96， F1 分數從 17.6%降到 12.8%。因為 DACSE2016 比賽主要看錯誤率，所以整體而言 DNN 方法還是明顯比 GMM 方法好。

關鍵詞：聲音事件偵測、深層類神經網路、音頻分析、多標籤分類

一、簡介

聲音是人類感知環境的重要資訊，也是反映人類行為的重要特徵。尤其是在某些環境中，一些特殊的聲音代表了某種狀況正在發生，例如：在辦公室裡，有鍵盤聲、開關門聲、笑聲、玻璃破碎聲…等，在居家環境中，有燒開水聲、嬰兒哭聲、跌倒聲、開門聲…等，或是在街頭環境下，有喇叭聲、碰撞聲、槍擊聲…等。

聲音事件偵測的實際應用很廣泛，例如：美國西雅圖政府日前公開展示一套槍聲偵測系統：ShotSpotter，用以更有效地遏止、打擊城市犯罪[1]。或是年老的長輩幾乎都獨自在家裡，在家中有可能會發生事情，例如：忘記自己正在燒開水，導致引發火災、在浴室跌倒，無法及時求救治療，頂樓窗戶被小偷打破，對家裡財物搜刮…等。此時若有音訊事件聲音偵測系統，就可以即時提供援助。

傳統的聲音事件技術主要可分為三個主流的技術類別，分別是以高斯混合模型(Gaussian Mixture Model, GMM)為基礎的語者辨識技術、以支持向量機(Support Vector Machine,SVM)為基礎的語者辨識技術，結合高斯混合模型與支持向量機(Hybrid GMM-SVM)之雙模型的語者辨識技術。

然而，在實際應用環境中，若遇到干擾偵測因素，例如：太多背景雜訊聲音的干擾或錄音品質太差等，傳統以高斯混合模型為基礎的語者辨識技術及以支持向量機為基礎的音訊辨識技術，因不具備環境適應的能力及對於錯誤容忍的程度太低，常會導致辨識系統的辨識性能無法維持。而對於結合高斯混合模型與支持向量機等兩者的模型而言，雖然該技術擷取兩類模型辨識技術的優點，但是其亦不具備環境適應與系統容錯的能力。這主要是因為 SVM 屬於淺層分析技術，因此訓練出來的模型，仍易受訊號的表面變易干擾[2][4]。

最近幾年，深層類神經網路被大量應用，因其可對訊號做深層分析，學習訊號的隱性結構，因此訓練出來的模型較不易受環境雜訊，不匹配的錄音設定…等等影響，具有強鍵性，可能較適合被應用到聲音事件偵測系統[3]。所以在論文中，我們將採用深層類神經網路，實做居家與戶外兩場景的音訊事件偵測系統，並據此參加 DCASE2016 評比，利用其具公信力的語料，尋找最佳的 DNN 設定。

二、相關研究

目前為止效能較高的聲音事件偵測模型大致分為下列幾種: (1) 傳統高斯混合模型(Gaussian Mixture Model, GMM)。 (2)支持向量機器(Support Vector Machine, SVM)。 (3) hybrid GMM/SVM。

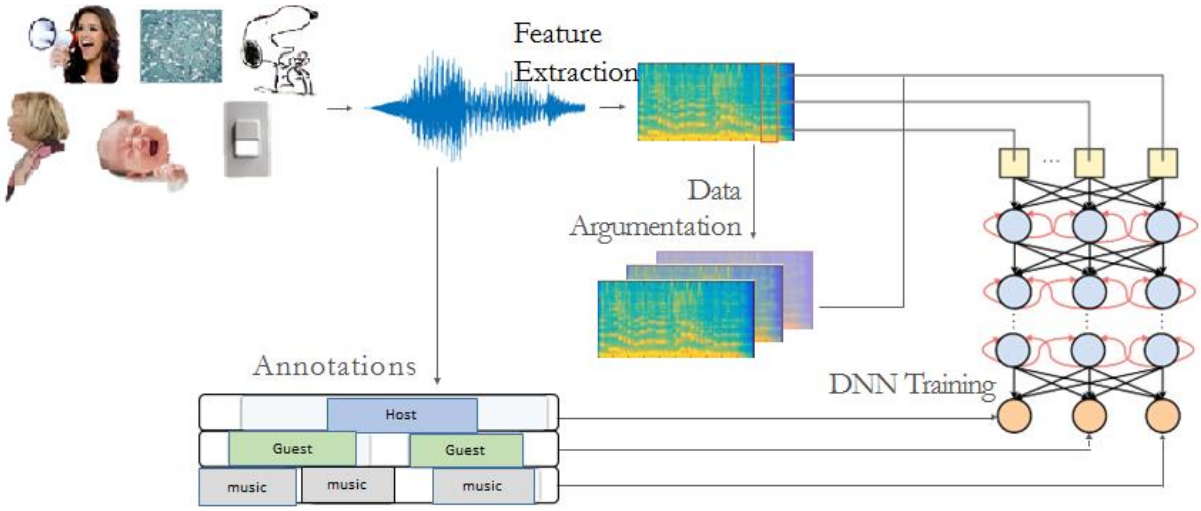
首先用高斯混合模型(GMM)來代表聲音模型的主要理由有兩個，第一個理由是高斯混合模型的每個基本密度函數可以模擬出一些聲音事件的特徵。因此我們可以用高斯混合模型中第 i 個平均值來代表第 i 個聲音特徵的頻譜形狀，而用共變異矩陣來代表頻譜形狀的變化。第二個理由是高斯混合模型能很平滑地近似任意形狀的密度。單一型態高斯混合語者模型是利用一個平均值向量和共變異矩陣來代表聲音事件特徵參數的分佈情形。而向量量化模型則是利用一組離散的特徵樣板來代表語者的分佈。高斯混合模型可以說是結合了上述兩種模型的優點，它利用了一組離散的高斯函數，加上高斯函數具有的平均值向量和共變異矩陣使得它有更好的模型能力。

此外，支持向量機器(SVM)的優勢在於使用上相當容易，SVM 主要要找出一個超平面(hyperplane)，使之將兩個不同的集合分開。以二維的例子來說，我們希望能找出一條分界線能夠將目標集合的樣本點和非目標集合的資料點分開，而且我們還希望這條分隔線距離這兩個集合的邊界(margin)越大越好，這樣我們才能夠很明確的分辨這個樣本點是屬於那個集合。

最後，Hybrid GMM/SVM 是高斯混合模型（GMM）和判別支持向量機（SVM）的結合。由於 SVM 模型和 GMM 模型各具優缺點，所以有研究提出建立 GMM 與 SVM 的混和模型，結合 GMM 對於數據表示特徵能力強與 SVM 對數據區分能力優良的特點，利用 GMM 對 SVM 的輸出做調整，實現 SVM 的概率輸出，以達到辨識率提升的目的。

三、深層類神經網路音訊事件辨認系統

在本實驗中我們使用 DNN 做音訊事件辨認系統，並用以參加 DCASE2016 比賽。DCASE2016 主辦單位提供的聲音事件資料分成環境與事件兩類，共有兩種環境與 18 種音訊事件，環境包括居家與戶外，其中居家環境中有 11 種音訊事件，戶外則有 7 種音訊事件[6]。圖一是我們使用的 DNN 音訊事件辨認系統架構。



圖一、DNN 音訊事件偵測系統

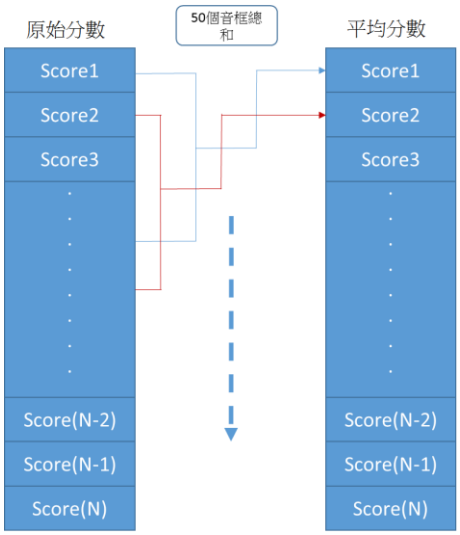
在此架構中我們使用多組 DNN 建立音訊事件偵測系統模型，因為在不同環境中，各種事件都有可能同時發生，所以每一種事件都需要建立一個獨立的 DNN 模型，平行做測試，主要為了確保當事件同時發生時，可以同時被系統偵測到[5]，DNN 模型訓練的示意圖如圖二。

圖二、DNN 訓練模型示意圖

其中建立各個音訊事件模型時，首先將訓練的音檔音框化之後，再擷取音訊事件的特徵參數(MFCCs)，我們先將所有訓練用音檔取梅爾倒頻譜參數，再個別收集各種事件本身

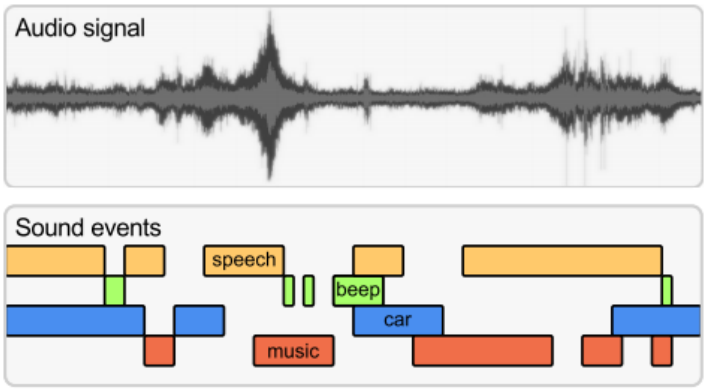
與非事件本身的音框用以訓練所有可能的 DNNs 模型[7]。

將所有事件模型訓練完畢之後，將測試音檔的音框個別送入各個事件模型，在測試時即可以得到各音檔的音框在不同事件時，為事件本身或不是事件本身的分數值，此外為求穩定判斷，我們再以 moving average 求取音框的平均分數當作最後的判斷依據，因此，最後分數的計算方式如圖三所示的分數計算示意圖。



圖三、音訊事件分數計算示意圖

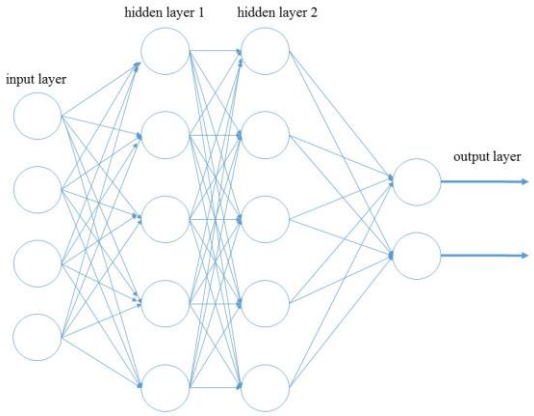
這次的 DCASE2016 挑戰[6]提供了十八種事件聲音，其事件與事件之間擁有同時發生的機會，故最後偵測的結果標籤需為多重標籤[3-5]，結果輸出規定的格式如圖四。



圖四、音訊事件偵測結果輸出方式

(一) DNN 原理

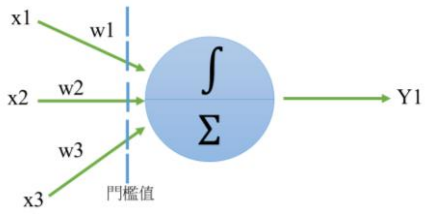
DNN 架構圖主要分為輸入層、隱藏層和輸出層其架構如圖五所示。輸入層在網路架構中為輸入訊息之一方，其神經元數目視輸入特徵參數數量而定。而隱藏層介於輸入與輸出層間，可以為複數層數，其使用非線性激活函數來萃取資訊，隱藏層中的神經元數量需要經由實際測試調整而定，隱藏層數量也跟神經元一樣都需經實驗獲得理想層數。最後，輸出層在網路架構中為提供資料輸出之一方，通常以一層表示，其神經元數目視輸出的內容而定。最後，深層神經網路通常具備至少二個以上的隱藏層，多出的隱藏層是為了提供更高的抽象層次，提高模型的能力。



圖五、多層 DNN 架構圖

DNN 中單一的神經元的運算方式如圖六所示，由輸入的參數 X 與連結權值 W，進行連乘加的動作，此一步驟可藉由集成函數(Summation Function)完成，集成函數的目的在於將前一層之輸出經由網路的連結權重值匯集至神經元中，通常是以函數的方式加以表達其公式如公式(1)，其中 W 為連結權值，X 為輸入變數，b 為該神經元的偏權值。經由此公式運算後，其輸出數值越大，則代表神經元被激發；輸出數值越小，則反之。最後再經由作用函數∫ (Activation Function)運算輸出，成為下一層神經元的輸入值。

$$f(x) = \int \left[\sum_{i=1}^{n} W_i X_i + b_j\right] \quad (1)$$



圖六、神經元運算方式示意圖

(二) DNN 訓練

由於在深層類神經網路中，所需調整的系統參數太多，因此 DNN 訓練通常使用 Gradient Descent，Gradient Descent 的公式如(2)表示。

$$x_t = x_t - \eta g_t \tag{2}$$

其中x_i為最佳化時要調整的參數，η是初始學習率，g_i為當前的梯度，此外在使用 Gradient Descent 演算法前，我們必須先定義一個 Cost Function 才能計算梯度，最常用的 Cost Function 為 cross entropy，其公式如公式(3)表示：

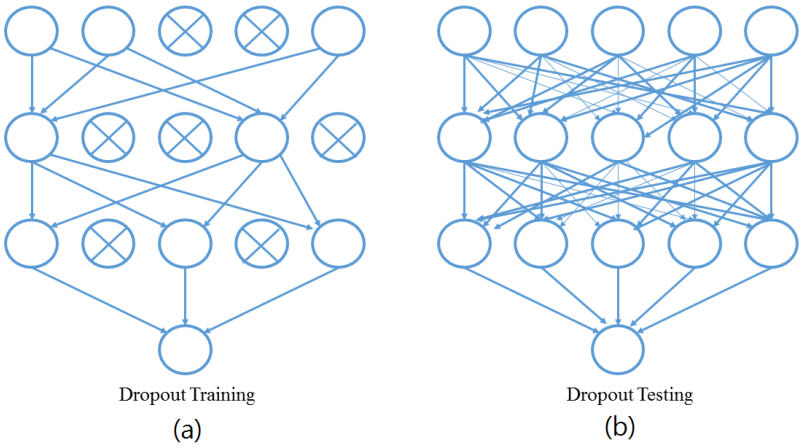
$$C= -\frac{1}{n}\sum_x [y ln_a + (1-y) ln(1-a)] \tag{3}$$

其中，a 為 DNN 輸出機率值，y 為正確類別指標答案。

(三) DNN 最佳化

此外為建立具強健性的 DNN 模型，常利用 Dropout 演算法。Dropout 在執行時，我們將隨機選擇忽略隱藏層萃取出的語料特徵，方式如圖七(a)所示，每個批次的 DNN 模型訓練過程中，因為每次隨機忽略的隱藏層語料特徵都不一樣，所以使每次模型中訓練到的類神經網路都將是不同的樣式，每次訓練都如同做一個新的音訊事件模型；除此之外，隱藏層語料特徵都是以一定機率隨機出現，因此不能保證每兩個隱藏層語料特徵每次都同時出現，這樣權重的更替不再依賴於有固定關係隱藏層語料特徵的共同作用，就不會出現某些特徵僅僅在其他特徵下才有效果的情況；但在測試語料輸入時，隱藏層特徵將不再隨機選擇忽略而是全部神經元的輸出的平均值如圖七(b)所示，如此可以使模型擁

有高抗雜訊能力[9]。



圖七、DNN 模型 Dropout 訓練示意圖

四、場景與音訊事件偵測實驗結果

實驗使用 DCASE2016 比賽提供的 TUT 資料庫，測試大會給予的 GMM baseline 與我們提出的 DNN 方法，DCASE2016 比賽類型分為 4 項，我們選擇其中的第三個任務，Sound event detection in real life audio(Task3)。Task3 任務是在評估我們日常生活中的音訊事件，其中的聲源都是在具有背景音干擾的情形下的多重事件。任務要求是當事件發生時，系統能否正確偵測到事件的發生，還有當多個事件同時發生時，系統是否能同時判斷出多個事件[8][10]。

(一)、實驗資料庫

DCASE2016 比賽提供的 TUT 語料中分為居家與戶外 2 個不同的場景，在不同場景各有不同的聲音事件，這些錄音在多個不同的位置錄製，包括不同的街道，不同的家庭。每次錄音時錄製一個 3~5 分鐘長，44.1 kHz 取樣率的音檔，每個音檔長度皆不同，事件長短也不一樣，並以人工依事件發生時間位置，給予標籤，當作標準答案。資料庫內容包含表一的訓練資料與表二的測試資料：

表一、DCASE2016，Task3 訓練資料庫

場景	聲音事件	音檔	聲音事件	音檔	場景	聲音事件	音檔
居家	(object) Rustling	41	Drawer	23	戶外	(object) Banging	15
	(object) Snapping	42	Glass jingling	26		Bird singing	162
	Cupboard	27	Object impact	155		Car passing by	74
	Cutlery	56	People walking	24		Children shouting	23
	Dishes	94	Washing dishes	60		People speaking	41
	Water tap running	37				People walking	32
						Wind blowing	22

表二、DCASE2016，Task3 測試資料庫

環境	音檔個數	音檔總長
居家	10	36min16s
戶外	12	42min

此外 TUT 資料庫用於記錄音訊之設備為 binaural sound engineer OKM II Kelaxike/studio A3 electret microphone Ear，並使用 44.1 kHz 採樣率和 24 位分辨率的 Roland Edirol R-09 waveform recorder 做錄音。

(二)實驗設定

我們先將所有訓練資料庫中音訊檔案取梅爾倒頻譜參數(MFCCs)，將其音框化之後，將每個事件的音框送進高斯混合模型訓練。最後使用測試資料庫中的音訊檔案進行測試時，先得到的每個音框的分數，再經過與閥值此對判斷，如分數到達標準，再將事件標記寫入文本，最後再拿文本與正確答案相互比對，得到錯誤率與 F1 分數，以下詳細說明各部分設定細節：

- 前處理：

求取梅爾倒頻譜參數時，其中濾波器數量為 40，梅爾倒頻譜參數為 20 維、頻率範圍取 0 Hz ~22050 Hz、傅立葉轉換為 2048，音檔使用的音框大小為 40ms。此外為了避免音框間的變化太劇烈，我們將兩個音框之間取 20ms 重疊。在實驗設定中不採用 MFCCs

的第 0 維，所以共 19 維 MFCCs，再加上一階與二階導數組成共 59 維特徵的向量，其中的一階與二階導數接考慮前後各四個音框。

- GMM 參數設定：

依據主辦單位給予的 GMM baseline 設定標準，在實驗當中，我們將每一個事件訓練為兩個模型，分別為事件本身、非事件的其他聲音。每個 GMM 模型使用混合數為 8，所以一個事件的 GMM 模型總混合數為 16(事件本身+非事件本身)。

- DNN 參數設定：

首先在一前置實驗中，我們先測試輸入音框數與類神經元數目，經把音框數設定為 1、5 或 9 做測試，結果在音框數為 1 時，擁有較佳結果。神經元數則曾經測試 32、64 或 128。結果在神經元數為 64、Dropout 試過 0.9、0.7、0.5，結果 0.7 時效果較好。因此，在以下實驗中皆採用輸入音框數為 1，神經元數 64 與 Dropout 為 0.7 的設定，並進一步測試 DNN 的層數。

(三)評估方式

系統的評分標準有兩種一種是錯誤率(error rate)，另一個則是 F1 分數。錯誤率的計算方式為：

$$ER = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} D(k) + \sum_{k=1}^{K} I(k)}{\sum_{k=1}^{K} N(k)} \tag{4}$$

其中的 N 代表正確答案的事件發生數。三組數據分別為：插入錯誤(Insertion,I)、取代錯誤(Substitution,S)及刪除錯誤(Deletion,D)。F1 分數的公式如下：

$$\text{F1}=\frac{2(Precision \cdot Recall)}{Precision+Recall} \tag{5}$$

其中的 precision 和 recall 為

$$\text{Precision} = \frac{tn}{tp+fp} \quad ，\text{Recall} = \frac{tp}{tp+fn} \tag{6}$$

此公式的符號定義如下：tp:正確判定為正確；fp:正確判定為錯誤；tn:錯誤判定為正確；

fn:錯誤判定為錯誤。

(四)實驗結果

首先測試音檔分為居家與戶外，我們分別使用 DCASE2016 大會給的 GMM 和我們提出的 DNN 模型做測試，在 DNN 系統中我們測試了使用一層和二層隱藏層的情況。音訊事件偵測分為兩大類實驗，共 3 個子實驗，兩大類包括(1)場景偵測和(2)居家與戶外音訊事件偵測。其中子實驗一為場景偵測，目的是要區分場景是在居家或戶外環境。子實驗二為居家環境音訊事件偵測，要在居家環境中要偵測 11 種音訊事件。子實驗三則為戶外音訊事件偵測，要偵測 7 種不同的音訊事件。

首先在子實驗一場景測試實驗方面，從表三的實驗結果來看，DNN 系統的平均總錯誤率 0.86，傳統模型 GMM 則為 0.91，DNN 的 F1 為 26.80%，GMM 的 F1 則為 23.40%，因此，以實驗結果來看，DNN 的錯誤率與 F1 都是最佳。但是以錯誤率來看，DNN 系統還有進步的空間，詳細結果如表三所示。

表三、Performance of Scene Recognition

	GMM		DNN			
#. of layers			1		2	
Scene	ER	F1	ER	F1	ER	F1
home	0.97	15.40%	0.93	13.20%	0.82	31.90%
residential	0.86	31.50%	0.95	11.50%	0.90	21.70%
Average	0.91	23.40%	0.94	12.30%	0.86	26.80%

表四為子實驗二居家音訊事件偵測的實驗結果。從平均錯誤率來看，GMM 為 1.06，而 DNN 則為 0.86，以 F1 分數來看，GMM 為 8.90%，而 DNN 則為 27.70%，總體而言在室內環境使用 DNN 音訊事件偵測系統比較好。

表四、室內環境音訊事件偵測錯誤率與 F1 分數

	GMM		DNN			
#. of layers			1		2	
Event	ER	F1	ER	F1	ER	F1
cupboard	1.00	0.00%	0.94	15.6%	0.93	22.00%
cutlery	1.02	0.00%	1.00	0.00%	0.56	62.80%
dishes	1.16	2.50%	0.98	3.70%	0.87	41.90%
drawer	1.19	0.00%	1.00	8.80%	0.92	26.00%
glass_jingling	1.10	0.00%	0.95	8.70%	0.70	54.00%
object_impact	1.06	19.30%	1.00	0.00%	0.99	1.50%
object_rustling	1.09	7.00%	1.00	0.00%	1.00	0.00%
object_snapping	1.00	0.00%	1.00	0.00%	1.00	0.00%
people_walking	1.10	14.80%	1.00	0.00%	1.00	0.00%
washing_dishes	1.08	20.30%	0.96	25.90%	0.92	29.70%
water_tap_running	0.83	34.10%	0.79	39.60%	0.54	66.70%
Average	1.06	8.90%	0.97	9.30%	0.86	27.70%

最後表五為子實驗三戶外環境的音訊事件偵測實驗結果。從平均錯誤率來看，GMM 為 1.03，而 DNN 則為 0.96，以 F1 分數來看，GMM 為 17.60%，而 DNN 則為 12.80%，雖然 DNN 的 F1 分數比 GMM 差，但是因這次比賽主要是比錯誤率，所以以整體來看，DNN 還是比 GMM 好。

表五、戶外環境音訊事件偵測錯誤率與 F1 分數

	GMM		DNN			
#. of layers			1		2	
Event	ER	F1	ER	F1	ER	F1
bird_singing	0.87	30.10%	1.04	3.60%	0.97	31.60%
car_passing_by	0.71	54.50%	0.77	37.70%	0.95	24.20%
children_shouting	1.07	0.00%	1.00	0.00%	1.00	0.00%
object_banging	1.00	0.00%	1.00	0.00%	0.82	34.00%
people_speaking	0.89	25.00%	1.00	0.00%	1.00	0.00%
people_walking	1.15	1.70%	1.00	0.00%	1.00	0.00%
wind_blowing	1.53	11.80%	1.01	2.20%	1.00	0.00%
Average	1.03	17.60%	0.97	6.20%	0.96	12.80%

五、結論

本研究使用 DNN，建立音訊事件聲學偵測系統。並利用 Dropout 達到最佳化 DNN 事件模型。降低通道雜訊干擾與背景環境的影響。實驗結果顯示 DNN 使用二層隱藏層，神經元數為 64 時，可在 DCASE2016 比賽測試資料中得到最佳結果。若與傳統 GMM 比較，其場景偵測錯誤率可從 0.91 降至 0.86、F1 分數並從 23.4%提升到 26.8%，此外針對室內環境的音訊事件偵測實驗，錯誤率可從 1.06 降至 0.86，F1 分數並從 8.9%提升到 27.7%，而對戶外環境的音訊事件偵測實驗，錯誤率可從 1.03 降至 0.96，F1 分數並從 17.6%到 12.8%，因為比賽主要是看錯誤率，所以從總結果來看，DNN 方法比 GMM 方法要好。所以我們提出的 DNN 架構確實是有效可行的。

致謝

本研究感謝教育部『大學以社教機構為基地之數位人文計畫』（A36 號）與科技部專題計畫（MOST 104-2221-E-027-079, 105-2221-E-027-119 and 103-2218-E-027-006-MY3）支持。

參考文獻

[1] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, A. Sarti, "SCREAM AND GUNSHOT DETECTION IN NOISY ENVIRONMENTS," in EURASIP European Signal Processing Conference (EUSIPCO 2007), Poland, Sept, 2007

[2] Emre Cakir, Toni Heittola, Heikki Huttunen and Tuomas Virtanen, "Polyphonic Sound Event Detection Using Multi Label Deep Neural Networks," in IEEE International Joint Conference on Neural Networks (IJCNN), 2015.

[3] Emre Cakir, Toni Heittola, Heikki Huttunen and Tuomas Virtanen, "MULTI-LABEL VS. COMBINED SINGLE-LABEL SOUND EVENT DETECTION WITH DEEP NEURAL NETWORKS ," 2015.

[4] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, “Contextdependent sound event detection,” in EURASIP Journal on Audio, Speech, and Music Processing, vol. 2013, no.

1, 2013, p. 1.

[5] Tara N. Sainath, Oriol Vinyals, Andrew Senior, Has¸im Sak, “Convolutional, long short-term memory, fully connected deep neural networks,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[6] Annamaria Mesaros, Toni Heittola, Tuomas Virtanen, "TUT Database for Acoustic Scene Classification and Sound Event Detection," in In 24rd European Signal Processing Conference 2016 (EUSIPCO 2016), 2016.

[7] Karol J. Piczak, “Environmental sound classification with convolutional neural networks,” in Proc. of MLSP). IEEE, 2015, pp. 1–6.

[8] D. Scherer, A. Muller, and S. Behnke. " Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition ". In ICANN. 2010

[9] Li Wan, Matthew Zeiler, Sixin Zhang, Yann LeCun, and Rob Fergus. "Regularization of NNs using DropConnect," In ICML, 2013.

[10] Grigorios Tsoumakas, Ioannis Katakis, “Multi-Label Classification: An Overview,” in Int J Data Warehousing and Mining, 2007, pp. 1–13